DOCUMENT RESUME

ED 315 439 TM 014 462

AUTHOR Campo, Stephanie F.
TITLE Alternative Logics for Estimating whether Research Results Will Generalize.
PUB DATE Nov 88
NOTE 18p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (17th, Louisville, KY, November 9-11, 1988).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Effect Size; *Estimation (Mathematics); Evaluation Methods; *Generalizability Theory; *Methods Research; *Sampling
IDENTIFIERS Cross Validation; *Efrons Bootstrap; *Jacknifing Technique; Tukeys Test

ABSTRACT

Three procedures for evaluating the sampling specificity of results are reviewed. These procedures are Tukey's jacknife technique, Efron's bootstrap technique, and cross-validation methods. The jacknife technique uses different subsamples derived from the original total data set to provide empirical estimates of the generalizability of effect sizes. The bootstrap technique estimates the statistical accuracy of effect size estimates and creates a megafile by copying the original data sample over and over again many times. The researcher then randomly selects a given number of bootstrap samples of size "n" from the megafile. The effect size is computed for each bootstrap sample; these correlation coefficients are treated as a distribution from which statistical estimates of result stability are derived. Cross-validation methods involve the arbitrary splitting of a sample. Prediction equations developed for each group are crossed so that each group will use the other group's prediction equations. A small data set is used to illustrate in more detail how the cross-validation procedure is performed and interpreted. Two data tables and sample Statistical Analysis System commands are provided. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED315439

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

STEPHANIE F. CAMPO

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Alternative Logics for Estimating Whether Research Results will Generalize

Stephanie Campo

University of New Orleans 70148

Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, Kentucky, November 9, 1988.

TM014462

ABSTRACT

Three procedures for evaluating the sampling specificity of results are reviewed: a) Tukey's jacknife technique, b) Efron's bootstrap technique, and c) cross-validation methods. Each procedure is briefly explained. A small data set is employed to illustrate in more detail how the cross-validation procedure is performed and interpreted.

Carver (1978) notes that "replication is the cornerstone of science" (p. 392) (Bauernfeind, 1968; Smith, 1970). The replication of research findings informs the researcher of the generalizability of obtained findings (Smith, 1970). Researchers need to know that observed effects are "true effects". Through investigation researchers endeavor to determine: a) the validity of sample-based results with respect to the broader population of interest; b) the stability of calculated findings derived from sample estimates; and c) the nature of the relationship between independent variables and observed phenomena.

Crask and Perreault (1977) point out that failure to determine the validity of sample-based results and the stability of calculated findings may lead to the reporting of inaccurate results based upon sample specific findings. The researcher's desire to demonstrate that results are replicable, i.e., that results are not based upon chance, has lead to a reliance upon statistical significance testing. Carver (1978) reports that interest in evaluating replicability is one of the two most influential reasons why statistical significance testing flourishes. Yet, the

interpretation of statistical significance as reflecting the probability that results will be replicated has no basis whatsoever in the logic of significance testing (Carver, 1978).

According to Carver, statistical significance testing limits the researcher to decisions of rejecting or failing to reject the null hypothesis given a probability of obtaining sample results under an assumption that the null hypothesis is exactly true. Therefore, the interpretation that the probability value reflects the replicability or reliability of results is completely erroneous.

Thompson (in press) has demonstrated through the use of example data how reliance upon statistical significance testing may mislead the researcher in the interpretation of results. Thompson employed varying sample sizes in illustrating that the value of the effect size remained unchanged even when the sample size increased; however, the interpretation of statistical significance did change as a function of the increase in sample size. The researcher basing decisions on statistical significance may ignore large effect sizes that are not significant while over interpreting effect

sizes that are small but statistically significant. Carver (1978) points out that effect sizes and significance testing do not inform the researcher of the likelihood of the replication of results. From a scientific point of view, it is more desirable to have a moderate effect size which is very stable or replicable rather than a large effect size which may be statistically significant but not stable or replicable.

There are three procedures for evaluating the sampling specificity of results: a) Tukey's jacknife technique, b) Efron's bootstrap technique, and c) cross-validation methods. This paper will describe how each procedure is performed and how results of each procedure are interpreted. A small data set developed by Thompson (in press) is employed to illustrate in more detail how the cross-validation procedure is performed and interpreted.

Jacknife Technique

The jacknife technique employes different subsamples derived from the original total data set to provide empirical estimates of the generalizability of effect sizes (Ayabe, 1985; Crask & Perreault, 1977). The stability of the jacknife estimate across subsamples

of the total data sample is interpreted by the researcher as an indicator of the reliability and replicability of the effect size obtained from the total sample.

The jacknife procedure is carried out by first computing the effect size for the entire sample, and then recomputing the statistic of interest n times, each time dropping a different observation from the sample (Ayabe, 1985). By repetitively dropping one observation at a time, the researcher is able to determine fluctutations in sampling error which may be attributed to the uniqueness of the single observation dropped or to the combined characteristics of the subsample. The standard deviations of estimated effect sizes derived with different subsamples indicate sampling error and enable the researcher to determine the stability of jacknife estimates. Crask and Perreault (1977) may be referred to for a readable presentation of the jacknife technique.

Bootstrap Technique

Like the jacknife technique, the bootstrap gives an estimate of the statistical accuracy of effect size estimates (Diaconis & Efron, 1983). However, in the

bootstrap technique, a megafile is created by copying the original data sample over and over again an extraordinary number of times. The researcher then randomly selects a given number of bootstrap samples of size n from the megafile. The effect size is computed for each bootstrap sample. These bootstrap correlation coefficients can be treated as a distribution from which statistical estimates of result stability may be derived (Diaconis & Efron, 1983).

The bootstrap technique is especially useful to the researcher when a large or moderate effect size is obtained, but a statistically nonsignificant finding has occurred due to a small sample size. In this case, the researcher can determine the replicability of results by performing the bootstrap. An example illustrating the application of the bootstrap is provided by Diaconis and Efron (1983).

Cross-validation

Cross-validation methods involve the arbitrary splitting of a sample. The sample may be split in half, or the sample may be split in other proportions such as sixty percent and forty percent. In cross-validation, the prediction equations developed for each of the split

groups are "crossed" so that each group will use the other group's prediction equations (Ayabe, 1985). The researcher wishes to determine two things: a) which beta weights (or related weights) will best predict the dependent variable from the predictor variables, and b) how stable in prediction is the effect size estimate. To make this discussion concrete, the use of cross-validation in regression research will be discussed in more detail.

To perform the cross-validation procedure, the researcher must carry-out two computer runs. The first run is conducted to derive the means and standard deviations for the total group and the two subgroups, which may also be referred to as invariance groups (Thompson, in press). The CORRELATION and MULTIPLE REGRESSION procedure are also run for the total group and for both invariance groups. The multiple correlation coefficient for the total group serves as the basis for ultimate interpretation if the results prove to be stable or invariant.

In the second computer run, the researcher creates new variables (i.e., **Z** scores and **YHAT** predicted scores). The **Z** scores for invariance groups use the

means and standard deviations for each of their respective groups. Two sets of **YHAT** values are created a) using their invariance group's data and beta weights and b) using their group's data and the other group's beta weights. Invariance results are obtained by running the CORRELATION procedure for all the **YHAT** values and dependent variable. Appendix A presents the SAS commands used to conduct the empirical analysis of Thompson's (in press) data set.

The researcher interprets the cross-validation results through a comparison of the multiple correlation coefficients, "shrunken" multiple correlation coefficients, and cross-validation or invariance correlation coefficients. In the cross-validation of results, the multiple correlation coefficients are obtained from a product-moment correlation between each subgroup's predicted scores derived using its own weights with the criterion group's scores (Krus & Fuller, 1982). The "shrunken $\underline{R}$" is obtained through the product-moment correlation between predicted scores of the subgroups derived using the other group's weights and the criterion scores. The cross-validation coefficient represents the product-moment correlations

between the predicted scores of the subgroups when each subgroup's own weights are applied as against when the other subgroup's prediction equation is applied.

While both the multiple correlation coefficient and the "shrunken R" represent correlations between the subgroup's predicted scores and criterion scores, the cross-validation or invariance coefficient represents the correlation between two sets of predicted scores. The researcher always hopes that the cross-validation or invariance coefficient will equal one.

The researcher looks for stability of the multiple R across subsamples and for the effects of "crossing" the regression equations for subgroups (i.e., "shrunken R"). If multiple R coefficients are comparable, then the researcher has some evidence for the replicability of results and for the representativeness of the sample (Krus & Fuller, 1982). However, the invariance coefficients can be directly interpreted, always against the standard of how close to one they are.

Table 1 presents a small data set for a multiple regression problem developed by Thompson (in press) . Two variables, "P" and "R" , are used to predict "DV", the dependent variable. The first three subjects were

randomly assigned to the first invariance subgroup ("INV" = "1"). The last four subjects were assigned to the second invariance subgroup ("INV" = "2").

INSERT TABLE 1 ABOUT HERE.

The invariance results produced by the CORRELATIONS procedure are presented in Table 2. The multiple correlation coefficients for the invariance groups are high, positive, and comparable; however, the "shrunken R" for invariance groups have a negative value which indicates that the regression equations are not generalizable across subgroups and therefore will not be generalizable to broader populations of interest. The invariance coefficients are also negative values which indicate a high degree of sampling error between subgroups. These data demonstrate that results are not replicable across subsamples.

INSERT TABLE 2 ABOUT HERE.

Thompson (in press) emphasizes the importance of empirically investigating result replicability rather than subjectively comparing the stability of multiple R

across subgroups. The cross-validation procedure will benefit the researcher wishing to demonstrate the replicability of results and the generalizability of sample characteristics.

Summary

Researchers need to know that observed effects are "true effects". Statistical significance testing does not inform the researcher regarding the important issue of whether results will generalize. The interpretation of statistical probability values as an indication of the likelihood that results will be replicated exceeds the logic of statistical reasoning. Further, reliance upon statistical significance testing may mislead the researcher in the interpretation of results. The researcher basing decisions on statistical significance may ignore large effect sizes that are not significant while over interpreting effect sizes that are small but statistically significant.

Three procedures for evaluating the sampling specificity of results are reviewed: a) Tukey's jacknife technique, b) Efron's bootstrap technique, and c) cross-validation methods. Each procedure is briefly explained. A small data set is employed to illustrate

in more detail how the cross-validation procedure is performed and interpreted.

References

Ayabe, Carol R. (1985). Multicrossvalidation and the jacknife in the estimation of shrinkage of the multiple coefficient of correlation. Educational and Psychological Measurement, 45, 445-451.

Carver, R.P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Crask, M.R., & Perrault, W.D., Jr. (1977). Validation of discriminant analysis in marketing research. Journal of Marketing Research, 14, 60-68.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. Scientific American, 248(5), 116-130.

Krus, D.J., & Fuller E.A. (1982). Computer assisted multicrossvalidation in regression analysis. Educational and Psychological Measurement, 42, 187-193.

Smith, N.C. (1970). Replication studies: A neglected aspect of psychological research. American Psychologist, 25, 970-975.

Thompson, B. (in press). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development.

Table 1
Observed and Latent Variables for Thompson's Data Set

INV	P	R	DV	ZP1	ZR1	YHAT11	YHAT12	ZP2	ZR2	YHAT21	YHAT22
1	1	3	90	-1	-.1323	.5151	-.87339	.	.	.	.
1	2	6	49	0	1.0595	-1.1525	.30350	.	.	.	.
1	3	1	93	1	-.9269	.6370	.57000	.	.	.	.
2	4	8	20	.	.	.	.	-1.162	.669	-.296	-.779
2	5	4	3	.	.	.	.	-.387	-.304	.474	-.411
2	6	0	39	.	.	.	.	.387	-1.276	1.245	-.042
2	7	9	63	.	.	.	.	1.162	.912	-1.423	1.232

Table 2
Invariance Statistics

	DV	YHAT11	YHAT12	YHAT21
YHAT11	1.0000[a] (n=3)			
YHAT12	-.2842[b] (n=3)	-.2843[c] (n=3)		
YHAT21	-.5182[b] (n=4)	.	.	
YHAT22	.8747[a] (n=4)	.	.	-.5924[c] (n=4)

[a] The multiple correlation coefficient R for the invariance group.

[b] The "shrunken R" for the invariance group.

[c] The invariance coefficient for the invariance group.

APPENDIX A: Example SAS Commands for Table 1 Data

```
DATA INVAR;
     INFILE INV;
     INPUT INV 1-2 P 4-5 R 7-8 DV 10-11;

if inv=1 then do;
     zp1=(p-2.0)/1.0;
     zr1=(r-3.333)/2.517;
     yhat11=(-.371189*zp1)+(-1.087694*zr1);
     yhat12=(.83549*zp1)+(.286434*zr1);
End;
Else Do;
     zp2=(p-5.5)/1.291;
     zr2=(r-5.25)/4.113;
     yhat21=(-.371189*zp2)+(-1.087694*zr2);
     yhat22=(.83549*zp2)+(.286434*zr2);
End;
PROC PRINT;
PROC MEANS;
     VAR P R DV;
PROC CORR;
     VAR P R DV;
     TITLE1 'DESCRIPTIVE STATISTICS FOR ALL DATA';
PROC REG;
     MODEL DV=P R/STB;
     TITLE1 'REGRESSION USING ALL DATA';


DATA TEMP1;
     SET INVAR;
     IF INVAR=1;
PROC CORR;
     VAR P R DV;
     TITLE1 'DESCRIPTIVE STATISTICS FOR SUBGROUP ONE';
PROC REG;
     MODEL DV=P R/STB;
     TITLE1 'REGRESSION FOR SUBGROUP ONE';


DATA TEMP2;
     SET INVAR;
     IF INVAR=2;
PROC CORR;
     VAR P R DV;
     TITLE1 'DESCRIPTIVE STATISTICS FOR SUBGROUP TWO';
```

```
PROC REG;
     MODEL DV=P R/STB;
     TITLE1 'REGRESSION FOR SUBGROUP TWO';


Data All;
     Set Invar;
Proc Corr;
     Var DV YHAT11 YHAT12 YHAT21 YHAT22;
     TITLE1 'INVARIANCE RESULTS';
```

Note. The analysis requires two runs. The first run includes procedures typed in all capitol letters. The second run procedures are typed in bold type. The newly created variables for the second run are typed in lower case.